ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 11 (2021) No. 3, pp. 1-26
DOI:10.33168/JSMS.2021.0301

Data Mining Methods and Techniques for Online Customer Review Analysis: A Literature Review

Anu Dahiya[1], Nidhi Gautam[1,*], Prashant Kumar Gautam[2]

[1] UIAMS, Panjab University

[2] UIHTM, Panjab University

anudahiya19@gmail.com; nidhig121@gmail.com (corresponding author); prashantgautam@pu.ac.in

Abstract. Data mining is an older term but is gaining importance in today's world. It is an art of extracting hidden information from the large data sets. Sometimes it can be described as a process of trawling for data to identify patterns which are previously unknown. Data mining tools can help to uncover the hidden knowledge in the large datasets and help in understanding customers in a better way. This paper reviews text mining basically online customer reviews. Customers' feedback in form of customer reviews plays an important role in forming opinion of potential customers who build their perceptions and make decisions by analyzing the facts and reviews of other customers. In this study, we conducted systematic literature review of online customer reviews covering background, trend and factors influencing the opinion of potential customers. We explored online customer's reviews by collecting, reviewing and synthesizing studies relating to online customer reviews published from 2015-2020. The result shows that studies related to online customer review analysis increased during the last 6 years. This review reveals online customer review analysis to be a promising new area of research. We highlight open problems and actual research trends. In the end, we conclude the paper by giving novel research trends in this field.

Keywords: Data mining, customer reviews, marketing, text mining.

1. Introduction

In present-day, big data has come out as a significant area that attracts the attention of various government, industries, academic and organizations worldwide (Mayer-Schönberger & Cukier, 2013, Thomson et al, 2014, Cuzzocrea, 2014). Various special issues published in the field of nature and science highlights opportunities and challenges of big data (Lynch, 2008, Science, 2011). Mckinsey, said that big data has pierced into or through every industrial sector. It becomes an important element in production (Manyika, 2011). Big data is used and mined for growth in production and consumer incitement. O'Reilly media declared that "the future belongs to the companies and people that turn data into products" (O'Neil & Schutt, 2013). For powering our future information economy, big data can be considered as the new fuel that will gear up our future information economy.

What big data means? Although, no universal definition of big data is coming out till now, but on the basis of a general agreement about its distinctiveness and uniqueness, it can be differentiated from traditional large data bases. Some researchers have proposed big data as a "moving definition" which changes overtime (Manyika et al, 2011). If the data is continuously increasing, it would be difficult to fix the threshold set to measure what type and size of data could be considered as big data. In a broader sense, physical world, human society as well as cyberspace can possibly be linked and coordinated through big data (Li & Cheng, 2012, Xiao-Long & Cheng, 2013, Cheng et al, 2014). So, big data can be divided into two heads, i.e., data from physical world and data from human society. Sensors, scientific experiments and observation can be considered to obtain data from physical world whereas data from human society is obtained from social networks, internet and other sources like finance, health, economics and transportation as well. Big data can be characterized by 5Vs, i.e., Volume, Velocity, Variety, Veracity and Value. Huge volume of data is not a challenge to cope up with, the real challenge lies around the diversity in large data sets, uncertainties in the data and to respond to the real time information (Sheng et al, 2017). For research perspective, big data includes structured data and unstructured data as well which is in the form of text, social media data, web data, multimedia as well as sensor data (Sheng et al, 2017). The social media marketing strategies plays a key role in the popularity of an organization. The Indian companies are extensively using the social media platform for as one of their marketing strategies but the number of companies using social media is still less and growing (Shin et al, 2020).

1.1. How Big the Data is?

Nowadays, organizations as well as individuals generate data in large quantity from which new insights can be obtained and competitive positions can be improved by using such insights. It is a general acceptance that organizational efficiency and competency can be enhanced using information obtained through big data

technologies. Until 2003, humans created 5 Exabytes (EB) of data which is now generated in just 2 days. According to CISCO report, devices as well as connections are growing more faster than growth in population and internet users worldwide. As per the report, former is growing at 10% CAGR while population and internet users are increasing at 1% CAGR and 7% CAGR respectively (Cisco VNI, 2018).

From socio-economic viewpoint, big data can be regarded as a supporter of second economy where activities of economic nature run on processors, sensors, executors and connectors. This concept of second economy was proposed by American economist W.B. Arthur in 2011 (Arthur, 2011).

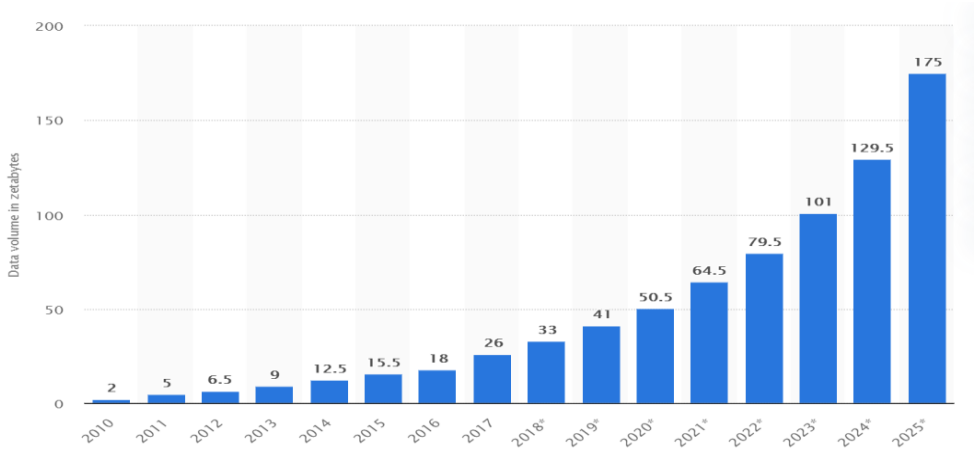

Fig. 1: Data volume from 2010 to 2025
Source: statista.com

Figure 1 shows the growth of big data from 2010 to 2025. Holst, 2019 suggests that amount of data in the world will increase dramatically in coming years and will reach at 175 Zettabytes by 2025. Furthermore, with the evolution of web 2.0 and implementation of IOT, more data will be generated (Holst, 2019).

1.2. Big Data Processing

There are basically two ways for big data pre-processing: **Machine learning** and **massively parallel processing.** Through machine learning techniques we can process large data in a most feasible way. It might help to classify big data into different categories, in understanding trends or movement of data, in detecting resemblance in data sets and predicting the future based on the past. Machine learning solutions on big data might help us in identifying fraud, bringing products faster in the market, and helps in becoming more and more competitive. Another way to process big data is massively parallel processing which includes massively parallel processing databases, data-mining grids, distributed file systems, distributed databases, cloud computing platforms and scalable storage systems (Chen et al, 2013).

For this study more focus would be on machine learning techniques. Both Machine Learning and Data Mining comes under the aegis of data science. Mostly same methods and techniques are employed in machine leaning and data mining and they often overlap each other significantly.

1.3. Data Mining

Data mining term is an older term but is gaining importance in today's world. It is an art of extracting hidden information from the large data sets. Sometimes it can be described as a process of trawling for data to identify patterns which are previously unknown. Secondary data is entirely concerned for analysis under data mining. In fact, data mining might be defined as the process of analyzing secondary large databases to identify patterns that might valuable for database owner (Hand, 1998).

In marketing, increased choices available for customers as well as intense competition make it difficult for marketing decision makers to take decisions quickly. It has created a new pressure on them because it would be complex to manage long term relationship with customers. But, to remain in the competition, firms need to maintain long-term relationship with their customer. Firms can manage customer relationship effectively only by understanding actual needs and wants of customers, rather than some assumed general characteristics (Meade, 1997, Peppers et al, 1999). For this, data mining tools can be used to discover the hidden information from the large datasets which further helps in understanding customers in a better way. Furthermore, knowledge management efforts are required in a systematic way to channelize knowledgeable information into effective marketing strategies. (Shaw et al, 2001). Therefore, in the current environment, where fulfilling customer's preferences are utmost important, need of simple as well as integrated framework for systematic management of customer knowledge is required. However, there are insufficient framework to create the link between management and customer knowledge extraction process. Studies on customer relationship have concentrated on various strategies required to manage customer interactions whereas studies on data mining have focused on the techniques. For true customer relationship management, integration of knowledge discovery process with management is required and it can be used further for marketing strategies. This will help marketers to address the needs and wants of customer on the basis of what marketer knows about customer's preferences, rather than generalized liking or disliking of customers.

As the market is becoming more and more competitive, organizations are looking for more valuable insights from the environment to better understand customer needs and wants. They are more focusing on increasing their number of customers and also finding ways to retain existing customers. Formerly, most of the data was in structured and semi-structured form and organizations were using it to retrieve meaningful information. But today's scenario is complex because 80% of the data generated worldwide is in unstructured form (Schneider, 2016).

Nowadays, it is easier for people to connect to the internet. As everything is online, people use internet to exchange their sentiments, opinion, emotions, etc. So, internet becomes integral part of human lives. With all the advancements in the technology and rise of social media there has been numerous platforms like blogs, social networks and discussion forms etc. where any individual can post his or her views more freely on various products, services and current issues.

As the cost of accessing internet is reducing, the easiness to access the web has been increasing significant. Customer find it convenient to shop online and make feedbacks online. When a customer wants to purchase product or avail any service from online websites whether it is booking hotels, purchasing products, booking cab, getting an insurance to everything, old customer's feedback in form of customer reviews plays an important role in forming opinion of those potential customers. They build their perceptions and make decisions by analyzing the facts and reviews of other customers.

Freedom of expression of customer views and opinions has given a new way to understand and analyze true and unbiased feedback. Even organizations are involved in finding true customer views so that they can use those reviews to get more insights from the textual data and use those insights for the betterment of product/ service. With the advent of Web 2.0, online customer reviews are widely available in real time. Firms can take advantage of easily accessible and real time information as they have very less time to react to the sudden change (Berger et al., 2010; Morinaga & Yamanishi, 2002). Furthermore, once any review is posted online, management of its diffusion becomes difficult as it is everlasting and can be retrieved at any time. Thus, analysis of opinion of customer over the time can be explored. These features of online user generated content might be the reason of increasing research interest in recent years (Chevalier & Mayzlin, 2006; Chintagunta, P. K., Gopinath, S., & Venkataraman, 2010; Duan et al., 2008; Godes et al., 2012; Li et al., 2008; Purnawirawan et al., 2014).

However, previous studies on customer-generated content have focused on aggregated measures like number of product reviews, star rating and average rating given by customers etc. but textual information which is a rich source of information were ignored. Now, researchers have started exploring the textual information in customer reviews and are interested to examine the effect of customer reviews on the attitude of customers, their intentions (Pavlou & Dimoka, 2006) as well as stock market performance of firms (Das et al., 2007; Sun, 2012). Researchers are also interested to explore customer reviews in order to extract the product attributes that are more relevant for customers' decision making (Archak, N., Ghose, A., & Ipeirotis, 2011; Decker & Trusov, 2010; Fowdur, L., Kadiyali, V., & Narayan, 2009), for analyzing the market structure (Lee & Bradlow, 2011; Morinaga & Yamanishi, 2002). There has been no effort to systematically review and synthesis studies on online customer review analysis to provide a clear idea of

online reviews to academicians as well as practitioners. Therefore, this study uses a systematic review protocol approach to explore online customer review analysis. To achieve the objective, we propose three key questions which are stated below:

1. What is the trend of publications related to online customer review analysis from 2015-2020?
2. What factors of online customer reviews influence the opinion of potential customers?
3. What are the limitations and gaps in current research on online customer reviews?

This systematic review of literature on online customer review analysis will contribute to the existing knowledge. Answering these questions may help the reader to know the trend in publication related to online customer reviews, factors influencing the opinion of potential customers and to highlight the limitations of previous studies and to identify gap.

2. Literature Review

2.1. Big Data

In the last decade, big data has been a buzzword. This term is coined by Roger Magoulas (Gantz & Reinsel, 2011). According to him, traditional data management tools are not sufficient to manage and process large data sets. Big data refers to diverse sets of large information that require complex computational platforms to analyze them (Akoka et al, 2017).

The term "Big Data" refers to providing right information using big data technologies to the right user at the right time from the large datasets growing exponentially from a long time. (Youssra & Sara, 2018).

Big Data refers to enormous datasets (Volume) which are diversified in various formats (Variety) such as structured, semi-structured, and unstructured data, and the data arrives faster than before (Velocity). Thus, there are 3Vs of big data, namely, Volume, Variety and Velocity.

There are two other "V"s which are important to be included in the definition of big data:

Veracity: it includes the degree of quality, accuracy as well as uncertainty in the data and sources of data.

Value: this refers to deriving value from the data that will be useful in business.

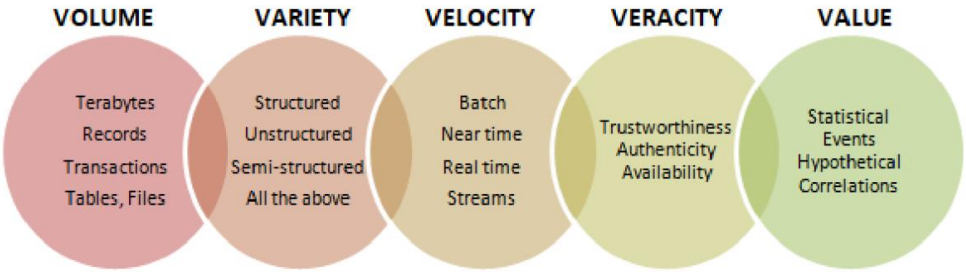

Fig. 2: 5Vs of big data
Source: Youssra & Sara, 2018

2.2. Data Mining

Data mining can be regarded as a new concept that lie at the interface of various disciplines such as machine learning, statistics, pattern recognition, database technology, and others. It is concerned with analysis of secondary datasets which are extremely large in size, and are of interest as well as value to the database owner (Hand, 1998).

Most of the statisticians are concerned with primary data analysis. First hand data is collected using experimental design and survey design for the particular questions' researcher have in their mind whereas data mining is entirely concerned with analyzing secondary data. Therefore, data mining techniques can be regarded as an inductive approach rather than deductive approach which is frequently seen as the paradigm for how modern science progresses.

2.2.1. Origin of Data Mining

Data mining stands at the junction of machine learning as well as statistics. There has been varied techniques in statistics like regression, discriminant analysis and principal component analysis for data exploration and model building. In classical statistics it is believed that the data are scarce and its computing is difficult. But in data mining, there is ample data and computing power. According to Daryl Pregibon, data mining can be described as "statistics at scale and speed" (Pregibon, 1999). This definition of data mining can be extended by adding "simplicity" to it. Simplicity refers not to the simplicity of algorithms but the simplicity in the logic of conclusion. Thus, data mining can be described as "statistics at scale, speed and simplicity". Furthermore, there are machine learning techniques, which are less structured as compared to statistical models and rely more on computational intensity. Decision trees and artificial neural network are some example of machine learning techniques.

2.2.2. Data Mining Functionalities

Table 1: Functionalities of Data Mining

Type of information	Explanation	Examples
Association	Relationship between set of items is linked in such a way that occurrence of one set of items is due to the presence of another set of items	Association rule can be used to identify which product can be sold together to a consumer
Cluster	Products with similar features can be grouped together	For market segmentation
Classification	Developing a class which can be used to decide the belongingness of certain item to that class	Fraud detection
Sequence	It involves events which occur in a particular sequence over an extended period	Customers ordering sheets might also buy comforter next
Similar time series	Refers to discovery of sequence similar to a given time	Finding stocks with a similar price movement
Exception	Anomaly detection	Detecting unusual credit card transactions
Forecasting	Estimation of future on the basis of certain known patterns in data	Estimating sales in future on the basis of previous records

Compiled from: Lei-Da Chen et al, 2000

2.2.3. Data Types

Traditionally we only recognize data in structured format, but now data is generated in semi-structured and unstructured format as well.

a. Data in a tabular format showing relationship between different rows and columns is known as structured data. They are easy to analyze as they stick to already defined data models.

b. Other format of data is Semi-structured data which falls between structured and unstructured data. These are self-describing data and may have simple label/value pairs. E.g. XML data, sensor data and JSON data. These data do not conform to a fixed structure. (Hurwitz et al, 2013).

c. Lastly, we have data in Unstructured format which does not have a pre-defined data models and specified format. It is typically text heavy data which is difficult to be compared with the data stored in structured databases. Nowadays only 20% is structured and the rest is in unstructured format. Text data, social media data, mobile data, videos, images are some examples of unstructured data.

As most of the data is in unstructured form i.e., around 80% of the data is

unstructured (Schneider, 2016), the current study focuses on unstructured data particularly text-based data. Now the study also shows that the video storytelling strategies are impacting the way consumers behave and purchase any product (Zou et al, 2021).

2.2.3.1 Textual Data and Text Mining

Most of the prior studies related to data mining have focused on structured data. But, now-a-days, substantial amount of data is available in text form. Data stored in text databases are growing rapidly. It contains various documents such as web pages, news articles, email messages, publications, world wide web, e-documents in electronic form (Konchady, 2006) and is stored in text databases of industries, businesses houses, government and various other institutes. This information is semi-structured in nature that means partially unstructured and partially structured (Berry, 2004). In research paper, a specific structured information is used such as title of the research paper, date of publication, authors etc. but, it also contains information in unstructured format like abstract, information about the topic, review of the existing studies (Gharehchopogh, 2010).

Various research on modeling and implementation of semi-structured data has already been done. Moreover, various information retrieval techniques, like text indexing methods have been developed to handle data in unstructured format. As unstructured data are plentiful, conventional information retrieval techniques become inadequate for its computation and management (Yin et al, 2007, Fan et al, 2006). Users need text mining tools for unstructured data to compare different documents, to identify relevant items of the documents as key words, and to discover various patterns and trends across multiple documents. Thus, text mining has become an important subset of data mining (Senellart & Blondel, 2003).

Text mining is similar to data mining (Senellart & Blondel, 2003), except that text mining can work with semi-structured and unstructured data and data mining tools (Inmon, 1996) are planned to handle only structured data. Thus, compared to data mining, text mining could be a better option.

Information available in text form can be divided into two heads, i.e., objective and subjective. Facts can be represented by objective statements while perceptions, perspectives or opinions are represented through subjective statements. In Natural Language Processing (NLP) the focus would be on mining factual information from the textual data. Thus, NLP is an example of objective statement. However, with the introduction of Web 2.0, some new and interesting ideas have been developed, which helps in extracting knowledge from user-generated content. Customer reviews in form of feedback can be obtained from various social networking sites. Answers to numerous research questions can be obtained by mining these reviews (Khan et al, 2014). Thus, with the introduction of Web 2.0, required information can be obtained from user generated content by applying information retrieval

techniques. This is known as subjective analysis. It is further subdivided into two domains, namely, opinion mining and sentiment analysis. According to few authors, opinion mining is similar to sentiment analysis (Liu, 2007), whereas others considered sentiment analysis as a sub-area of opinion mining (Tang et al, 2009). There is a little difference in opinion mining and sentiment analysis according to Morinaga et.al, 2002. In sentiment analysis, classification of a text in either positive or negative whereas information retrieval, analysis of rating given by user to represent his opinion is related to opinion mining (Morinaga et.al, 2002). According to Pang et. al (2008), opinion mining is extremely useful in practical applications, even though it is an intellectually difficult problem.

2.2.3.2 Text Data, Text Mining and Marketing

In a broad sense, marketing is "the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large" (AMA, 2019). Selling of particular products, service and brand can be promoted by having efficient communication between firms and customers. Through this lens, it is required to understand consumers for making better marketing decisions. Marketing is a seamless amalgamation of science and art. The art part deals with product design (looks), advertisement and maintaining personal relationships with customers, while the scientific part deals with including features within products, customer segmentation and analyzing customer feedback etc. Knowledge of customers' choice and preferences are to be kept in mind while doing the later part. However, during the production era when production was given highest level of importance, customer needs were not considered important. But today's scenario is different, markets are becoming more competitive and the customers are having more options while purchasing any goods. Customers started becoming more powerful, in fact, powerful enough to dictate the market. Customer needs are given utmost importance.

Marketers are now able to understand that for sustaining in the market, it is important to connect to customers with their product. Hence, understanding customer need became an indispensable task for marketers. However, it is not easy to gather market information. It requires a structured methodology with strong foundations. Formerly, most of the data was in structured and semi-structured form and organizations were using that data to retrieve meaningful information & that was quite simple. But today's scenario is different. It is estimated that 80% of the world's data is in unstructured form- text, image, audio and video and it becomes quite complex for organizations to extract meaningful insights from the data that is text-heavy (Schneider, 2016). Nowadays, it is easier for people to connect to the internet. As everything is online, people use internet to exchange their sentiments, opinion, emotions, etc. So, internet becomes integral part of human lives.

Market research can be done in many ways. But, broadly speaking, it can be of

two types, i.e., qualitative research and quantitative research. In quantitative research, usually, a structured questionnaire is built and very specific hypotheses are tested on the basis of the responses from the respondents against the questionnaire. Since, real-life data are also contained within unstructured texts, text mining approach can be used to extract business relevant information from such textual data in a cost effective and speedy way. If we look at the impact of internet on our lives, it is to be understood that internet has changed the way we interact with each other. Consumer reviews can act as a very effective source of information in today's world and mining those reviews could provide important piece of information from business contexts very easily.

2.2.4. Online Consumer Feedback and Decision Making

According to a survey report on over 5500 online customers, expert reviews are not considered valuable as compared to customer reviews by 59% respondents. Consumer generated reviews are given more importance (Gan et al, 2017). With the advent of Web 2.0, huge amount of data is generated and posted on the web. Feedback in form of online customer reviews are growing rapidly. It includes customers' thoughts, opinions and their experiences associated with brands (Archak et al, 2011, Decker & Trusov, 2010, Morinaga et al, 2002, Zhu & Zhang, 2010). This will help managers to understand preferences of customers and to track and monitor online comments of their product or brand (Gensler et al, 2015).

Online product reviews are an alternative way to collect data. It provides useful information to managers. Online product reviews affect purchase behavior of customers (Decker & Trusov, 2010, Baek et al, 2012, Li et al, 2013, Tirunillai & Tellis, 2012) and firm's performance as well (Floyd et al, 2014, King et al, 2014). So, managers may extract such valuable insights from online reviews and act accordingly. Consumer generated content especially online customer reviews on various brands is a rich source of information as it lacks biasness. It is considered as a friendly suggestion among customers (East et al, 2008).

With the advent of Web 2.0, online customer reviews are widely available in real time. Firms can take advantage of easily accessible and real time information as they have very less time to react to the sudden changes (Morinaga et al, 2002, Berger et al, 2010). Furthermore, once any review is posted online, management of its diffusion becomes difficult as it is everlasting and can be retrieved at any time. Thus, analysis of opinion of customer over the time can be explored. These features of online user generated content might be the reason of increasing research interest in recent years (Chevalier & Mayzlin, 2006, Chintagunta et al, 2010, Duan et al, 2008, Godes & Silva, 2012, Li & Hitt, 2008, Purnawirawan et al, 2012). However, previous studies on customer-generated content have focused on aggregated measures like number of product reviews, star rating and average rating given by customers etc. but textual information which is a rich source of information were ignored. Now, researchers have started exploring the textual information in

customer reviews and are interested to examine the effect of customer reviews on the attitude of customers, their intentions (Pavlou & Dimoka, 2006) as well as stock market performance of firms (Das & Chen, 2007, Sun, 2012). Researchers are also interested to explore customer reviews in order to extract the product attributes that are more relevant for customers' decision making (Archak et al, 2011, Decker & Trusov, 2010, Fowdur et al, 2009), for analyzing the market structure (Morinaga et al, 2002, Lee & Bradlow, 2011).

Online customer reviews in text format are qualitative in nature and star ratings are considered as quantitative information. Qualitative information complements quantitative information as the reviewers have to justify the reason why they have rating a product the way they did. Thus, potential buyers become more prudent in making interpretations of the information provided by purchasers in form of online reviews (Kumar & Benbasat, 2006). Firms can use online customer reviews as a tool for gaining confidence of customers. According to Kumar & et. al, 2006, as online customer reviews are easily available on websites, online retail websites become more useful. However, reviews having stronger effects on customer purchase decisions are considered more helpful by consumers (Chen et al, 2008, Chevalier & Mayzlin, 2006). Additionally, McKnight & et.al, 2006 indicated that credibility of information is the most important factor in e-WOM adoption. Thus, credible consumer reviews should be provided by online retail market to pursue success.

Kumar et.al, 2016 in their study concentrated on mining customer reviews from Amazom.com. This is the platform where consumers freely provide reviews relating to product. Researchers have used 3 classifiers to classify reviews as positive or negative. Results shows that Naïve Bayes classifier proves to be the most efficient in classifying reviews as positive or negative as compared to Logistic Regression and SentiWordNet.

Gan et.al, 2017 in their study wants to identify the structure of online customer reviews of restaurants and bars of Phoenix city in Arizona, U.S. listed on Yelp.com. They have also examined the effect of sentiments and review attributes on star ratings of restaurant. For this, they have used Yelp data base which consists of 335,022 customer reviews 15,585 businesses. After analysis, results show that food, shelter and context have more effect on star ratings as compared to price and ambiance.

Table 2: Summary of review of literature of online customer review analysis

Authors	Findings
Konchady, 2006	Most of the prior studies related to data mining have focused on structured data. But, now-a-days, substantial amount of data is available in text form. Data stored in text databases are growing rapidly.
Yin et al, 2007; Fan et al, 2006	Unstructured data are plentiful, conventional information retrieval techniques become inadequate for its computation and management
Schneider, 2016	80% of the world's data is in unstructured form
Gan et al, 2017	According to a survey report on over 5500 online customers, expert reviews are not considered valuable as compared to customer reviews by 59% respondents.
Gensler et al, 2015	Online customer reviews will help managers to understand preferences of customers and to track and monitor online comments of their product or brand.
Decker & Trusov, 2010; Baek et al, 2012; Li et al, 2013; Tirunillai & Tellis, 2012; Floyd et al, 2014; King et al, 2014	Online product reviews are an alternative way to collect data. It provides useful information to managers. Online product reviews affect purchase behavior of customers and firm's performance as well.
Morinaga et al, 2002; Berger et al, 2010	Firms can take advantage of easily accessible and real time information as they have very less time to react to the sudden changes
Chen et al, 2008; Chevalier & Mayzlin, 2006	Reviews having stronger effects on customer purchase decisions are considered more helpful by consumers.

3. Systematic Review Process

Systematic literature review is a process where each step builds upon the previous step. The process starts with a research question and ends on results and reports. The following are the stages of systematic review for extraction of review papers related to online customer reviews. We have started with identifying the 3 key research questions which will be answered by following a search strategy.

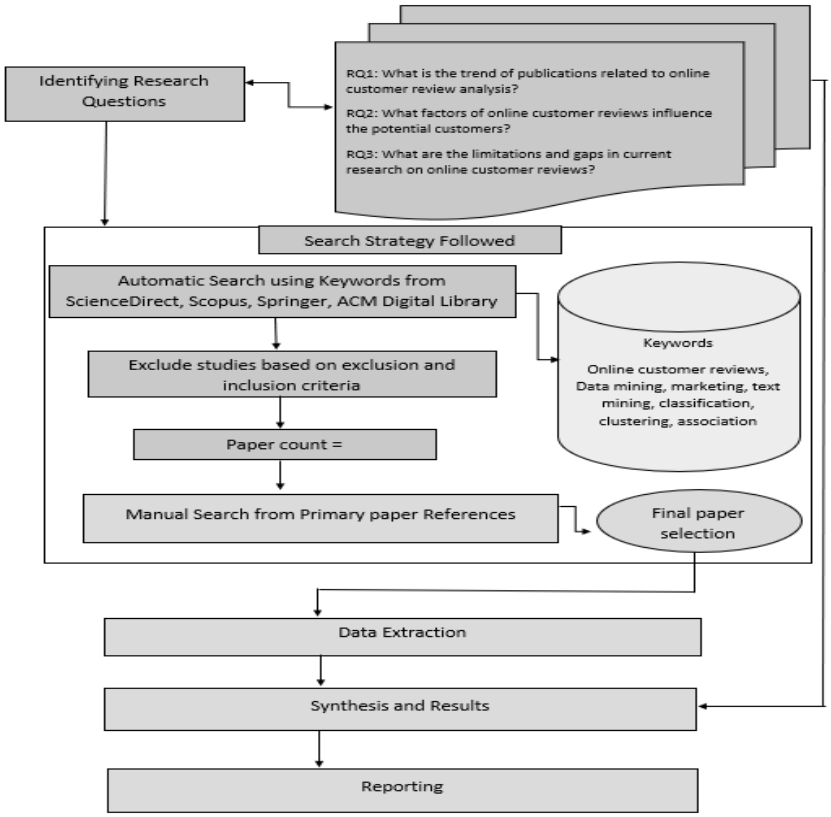

Fig. 3: Stages of systematic literature review

3.1. Conducting Research

A search strategy was followed wherein research papers from ScienceDirect, Scopus, Springer and ACM Digital Library were collected using keywords- online customer reviews, data mining, marketing, text mining, classification, clustering and association. Inclusion and exclusion criteria are used to exclude studies that are not relevant to answer the specific research questions. We have included every paper from the base corpus until excluded. Studies which are of no use and are not relevant to answer current research questions of the study are excluded.

3.2. Screening of Papers for Inclusion and Exclusion

Keywords mentioned in Fig. 3 were used to search research papers for current study. Every paper is included from the base corpus until some exclusions are made. Using the keywords 70 papers has randomly been included. Now exclusion criteria is followed wherein a research paper has gone under 3 stages for exclusion. The first exclusion criteria is “irrelevant abstract”, the second criteria is “abstract do not clearly defined contribution of work” and the third criteria is “abstract clearly not related to online customer reviews”

Table 3: Exclusion criteria

No.	Exclusion criteria
1	Irrelevant abstract
2	Abstract do not clearly defined contribution of work
3	Abstract clearly not related to online customer reviews

After screening of papers, 42 research papers were studied for answering the research questions of our study.

4. Findings

4.1. Trends of Publications Related to Online Customer Reviews from 2015-2020

Table 4: Publication trend for online customer reviews from 2015-2020

Author(s)	Topic of the study	Year	Findings
Liu, Z., & Park, S. (2015)	What makes a useful online review? Implication for travel product websites.	2015	Review messages' qualitative characteristics (i.e., perceived enjoyment and readability) make greater contributions to explaining the review usefulness beyond the other characteristics, such as messengers' and reviews' quantitative factors.
Askalidis & Malthouse, 2016)	The Value of Online Customer Reviews	2016	people usually don't pay attention to the entire set of reviews, especially if there are a lot of them, but instead they focus on the first few available.
Su J.K., Ewa M. & Edward C. M. (2017)	Understanding the effects of different review features on purchase probability	2017	Argument quality (measured as review length) has an inverted-U shaped relationship with purchase probability.
Helversen et al., 2018	Influence of consumer reviews on online purchasing decisions in older and younger adults	2018	Older adults were strongly influenced by affect-rich negative consumer reviews, but not by better average consumer ratings or affect-rich positive reviews Whereas youngers were strongly influenced by average consumer ratings and positive affect-rich reviews
Chen et al., 2019	Measuring and Managing the Externality of	2019	managerial responses can significantly influence subsequent customer review behavior. As such, managerial

	Managerial Responses to Online Customer Reviews		responses could be a valuable tool for businesses to interact with customers in the online world.
Jaiswal & Singh, 2020	Influence of the Determinants of Online Customer Experience on Online Customer Satisfaction	2020	economic value, customization, post-purchase experience and customer services are the major factors on which customers evaluate their overall online experience and satisfaction.

4.2. Factor of Online Customer Reviews Influencing the Opinion of Potential Customers

Table 5: Customer review KPIs for potential customers

Author(s)	Method	Predictors/ Factors	Outcomes	Findings
Baber et al. (2016)	Survey	Trustworthiness, Expertise, Experience, WOM use	Intension to purchase electronic products.	e-WOM formed the level of trustworthiness among customers and positively affected attitude which has an impact on purchase intensions
Cheng and Ho (2015)	Secondary analysis of exiting reviews	Argument quality, Source credibility	Review usefulness	Argument quality and source credibility all have a significant positive effect on the reader's perception of the usefulness of reviews. The effects of source credibility have more effect than argument quality
Kim, and Seo (2015)	Experiment	Perceived authority, Perceived bandwagon, Perceived objectivity, social plugins, Star ratings	Product attitude Webpage attitude Purchase intention	Expert reviews have a greater impact on attitudes toward product review websites, and this effect was moderated by star ratings. Star ratings have a strong positive effect on users' attitudes toward the product, attitudes toward the website, and their purchase intention.
Park, and Han (2007)	Experiment	Review quantity, Review quality	Purchase intension	The quality and quantity of online reviews positively affect consumers' purchase intention. Low-involvement consumers are affected by the quantity rather than the quality of reviews, whereas high-involvement consumers are affected by review quantity mainly when the review

				quality is high
Zhang et al. (2014)	Survey	Argument quality, Source credibility, Perceived quantity of reviews	Behavioral intensions	Argument quality, source credibility, and perceived quantity of reviews were key determinants of behavioral intention.
Park and Kim (2008)	Experiment	Expertise of review readers, Types of reviews, Number of reviews	Purchase intensions	The effect of type of reviews (cognitive fit) on purchase intention is stronger for experts than for novices while the effect of the number of reviews on purchase intention is stronger for novices

Compiled from: Su, Ewa & Edward (2017)

4.2.1. Word Cloud of Factors of Online Customer Reviews Influencing Opinion of Potential Customers

Fig. 4: Factors of online customer reviews influencing opinion of potential customers

Table 6: Frequency of words

Word	Frequency
Reviews	7
Effect	6
Quality	6
Intention	5
Purchase	5
quantity	4

affected	3
argument	3
attitudes	3
consumers	3

There are factors of online customer reviews which influence the opinion of potential customers. Existing literature reveals that high-involvement consumers are affected by review quantity mainly when the review quality is high. Star ratings have a strong positive effect on users' attitudes toward the product, attitudes toward the website, and their purchase intention. Moreover, the effects of source credibility have more effect than argument quality.

4.3. Research Challenges and Future Research Directions

Previous studies on online customer review analysis were carried out worldwide with skewed focus on factors which satisfy or dissatisfy a customer. For this, most of the studies have used quantitative information i.e., star ratings. Qualitative information in form of text provides rich information about the purchase experience of customers and very few studies have worked on it. There are studies in literature which have used longitudinal data to measure the changes over time in the effectiveness of review helpfulness. Number of reviews, star ratings and volume of review were given more importance rather than the content in the review.

Number of studies have worked on theory-driven approaches to suggest attributes of online customer reviews. Researchers can work on data-driven approaches for deriving such attributes of online reviews. Classification algorithms were used to analyze their efficiency for text classification in many studies. Researchers can also use various classification algorithms in a single study to decide the best text classifier. Moreover, most of these studies have been mainly undertaken to understand the consumer's response towards a product using various data mining techniques. Researchers used classification algorithms, clustering algorithms or making association rule or sequence patterns separately or all together in a single study to predict future behavior of customers based on online reviews.

There are no or very few studies that compare the quantitative and qualitative information of online customer reviews. Text-based reviews are qualitative in nature and star ratings are quantitative. Sometimes qualitative and quantitative information are contradictory in case of online customer reviews.

Most of the studies are limited to a single industry or cover limited geographical regions, particularly developed countries. In future, more industries can be covered in a single study and comparative analysis of geographical area can be done.

Moreover, existing studies have used various algorithms to analyze online reviews but did not mention the sequence of data mining techniques to follow which may result in more detailed explanation of review to analyze the behavior of

potential customers.

5. Conclusion

This paper focused on the data mining particularly mining online customer reviews which are in the form of text. We have shown the recent trends in the publication of online customer reviews and the factors of online reviews that majorly influence the opinion of potential customers. Study reveals that high-involvement consumers are affected by review quantity mainly when the review quality is high and star ratings have a strong positive effect on users' attitudes toward the product, attitudes toward the website, and their purchase intention. Moreover, the effects of source credibility have more effect than argument quality.

Economic value, customization, post-purchase experience and customer services are the major factors on which customers evaluate their overall online experience and satisfaction.

This research paper has also addressed the current challenges and future directions. Future research opportunities are abundant in this field as 80% of the data generated is in unstructured form, so it is important to extract useful insights from the unstructured data.

References

Akoka, J., Comyn-Wattiau, I., & Laoufi, N. (2017). Research on Big Data–A systematic mapping study. *Computer Standards & Interfaces*, 54, 105-115.

Askalidis, G & Malthouse, E.C. (2016). The Value of Online Customer Reviews. *ACM International Conference proceeding series*, 15-19.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management science*, 57(8), 1485-1509.

Arthur, W.B. (2011). The second economy, available at: http://www. images-et-reseaux. com/sites/default/files/medias/blog/2011/12/the-2nd-economy.pdf, 2011.

Baber, A., R. Thurasamy, M.I. Malik, B. Sadiq, S. Islam, and M. Sajjad (2016). Online word-of-mouth antecedents, attitude and intention-to-purchase electronic products in Pakistan. *Telematics and Informatics*, 33(2), 388-400.

Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of online consumer reviews: Readers' objectives and review cues. *International Journal of Electronic Commerce*, 17(2), 99-126.

Berger, J., Sorensen, A. T., & Rasmussen, S. J. (2010). Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, 29(5), 815-827.

Berry, M. W. (2004). Survey of text mining. *Computing Reviews*, *45*(9), 548.

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: a data management perspective. *Frontiers of Computer Science*, *7*(2), 157-164.

Chen, P. Y., Dhanasobhon, S., & Smith, M. D. (2008). All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon. com. Com (May 2008).

Wei C., Bin G., Ye Q., Kevin, Zhu X. (2019). Measuring and Managing the Externality of Managerial Responses to Online Customer Reviews. *Information Systems Research. Articles in Advance*, 1-16

Cheng, Y.-H., & H.-Y. Ho. (2015). Social influence's impact on reader perceptions of online reviews. *Journal of Business Research*, 68(4), 883-7.

Cheng, X. Q., Jin, X. L., Wang, Y., Guo, J., Zhang, T., & Li, G. (2014). Survey on big data system and analytic technology. *Journal of software*, *25*(9), 1889-1908.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book s. *Journal of marketing research*, 43(3), 345-354.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing science*, 29(5), 944-957.

Cuzzocrea, A. (2014, November). Privacy and security of big data: current challenges and future research perspectives. In *Proceedings of the first international workshop on privacy and secuirty of big data*, 45-47.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science*, 53(9), 1375-1388.

Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 27(4), 293-307.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. *Decision support systems*, 45(4), 1007-1016.

East, R., Hammond, K., & Lomax, W. (2008). Measuring the impact of positive and negative word of mouth on brand purchase probability. *International journal of research in marketing*, 25(3), 215-224.

Enterprise big data framework retrieved from https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/ on 26 Dec, 2019

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76-82.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing*, 90(2), 217-232.

Forsey, C. (2018). The Top 7 Search Engines, Ranked by Popularity retrieved from https://blog.hubspot.com/marketing/top-search-engines on 1 February, 2020

Fowdur, L., Kadiyali, V., & Narayan, V. (2009). The impact of emotional product attributes on consumer demand: An application to the US motion picture industry. *Johnson School Research Paper Series*, 22-09.

Gan, Q., Ferns, B. H., Yu, Y., & Jin, L. (2017). A text mining and multidimensional sentiment analysis of online restaurant reviews. *Journal of Quality Assurance in Hospitality & Tourism*, *18*(4), 465-492.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iview*, *1142*(2011), 1-12.

Gensler, S., Völckner, F., Egger, M., Fischbach, K., & Schoder, D. (2015). Listen to your customers: Insights into brand image using online consumer-generated product reviews. *International Journal of Electronic Commerce*, 20(1), 112-141.

Gharehchopogh, F. S. (2010, October). Approach and review of user oriented interactive data mining. In *2010 4th International Conference on Application of Information and Communication Technologies*, 1-4. IEEE.

Godes, D., & Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3), 448-473.

Hand, D. J. (1998). Data mining: statistics and more? *The American Statistician*, *52*(2), 112-118.

Helversen, B.V et. al. (2018). Influence of consumer reviews on online purchasing decisions in older andyounger adults. *Decision Support Systems*, 1-10, 113.

Holst, A. (2019). Data volume from 2010 to 2025 retrieved from https://www.statista.com/statistics/871513/worldwide-data-created/ on 19 December, 2019

https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html#_Toc532256792 retrieved on 28 November, 2019.

https://en.wikipedia.org/wiki/Machine_learning retrieved on 24 December, 2019.

https://www.ama.org/the-definition-of-marketing-what-is-marketing/ retrieved on 30 Dec,2020.

Hurwitz, J. S., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big data for dummies*. John Wiley & Sons.

Inmon, W. H. (1996). The data warehouse and data mining. *Communications of the ACM*, *39*(11), 49-51.

Jaiswal & Singh. (2020). Influence of the Determinants of Online Customer Experience on Online Customer Satisfaction. *Sage Publications.* 24(1) 41–55.

Khan, K., Baharudin, B., Khan, A., & Ullah, A. (2014). Mining opinion components from unstructured reviews: A review. *Journal of King Saud University-Computer and Information Sciences*, 26(3), 258-275.

King, R. A., Racherla, P., & Bush, V. D. (2014). What we know and don't know about online word-of-mouth: A review and synthesis of the literature. *Journal of interactive marketing*, 28(3), 167-183.

Kim, H.-S., P. Brubaker, and K. Seo. 2015. Examining psychological effects of source cues and social plugins on a product review website. *Computers in Human Behavior*, 49, 74–85.

Konchady, M. (2006). *Text mining application programming*. Charles River Media, Inc..

Kumar, K. S., Desai, J., & Majumdar, J. (2016, December). Opinion mining and sentiment analysis on online customer review. *In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 1-4. IEEE.

Kumar, N., & Benbasat, I. (2006). Research note: the influence of recommendations and consumer reviews on evaluations of websites. *Information Systems Research*, 17(4), 425-439.

Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5), 881-894.

Lei-da Chen, T. S., & Frolick, M. N. (2000). Data mining methods, applications, and tools. *Information systems management*, 17(1), 67-68.

Li, G. J., & Cheng, XQ. (2012). Research status and scientific thinking of big data. *Bull Chin Acad Sci*, 27, 647–657 (in Chinese)

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4), 456-474.

Li, M., Huang, L., Tan, C. H., & Wei, K. K. (2013). Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, 17(4), 101-136.

Liu, B. (2007). Web data mining: exploring hyperlinks, contents, and usage data. *Springer Science & Business Media*.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, 47, 140-151.

Lynch, C. (2008). How do your data grow? *Nature*, *455*(7209), 1-136.

Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. *http://www. mckinsey. com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011, May). Big Data: The Next Frontier for Innovation, Competition, and Productivity. *McKinsey Global Institute, San Francisco*.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McKnight, H., & Kacmar, C. (2006, January). Factors of information credibility for an internet advice site. *In Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, 6, 113b-113b.

Meade, B. (1997). Enterprise One to One: Tools for Competing in the Interactive Age. *Journal of Business and Entrepreneurship*, *9*(2), 73.

Morinaga, S., Yamanishi, K., Tateishi, K., & Fukushima, T. (2002, July). Mining product reputations on the web. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 341-349.

O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1-2), 1-135.

Park, D.-H., J. Lee, and I. Han. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, 11(4), 125-48.

Park, D.-H., and S. Kim. 2008. The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews. *Electronic Commerce Research and Applications*, 7(4), 399–410.

Pavlou, P. A., & Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research*, 17(4), 392-414.

Peppers, D., Rogers, M., & Dorf, B. (1999). Is your company ready for one-to-one marketing? *Harvard business review*, *77*(1), 151-160.

Purnawirawan, N., Dens, N., & De Pelsmacker, P. (2012). Balance and sequence in online reviews: The wrap effect. *International Journal of Electronic Commerce*, 17(2), 71-98.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision support systems*, *31*(1), 127-137.

Schneider, C. (2016). https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/ retreived on 26 Dec, 2019

Science. (2011). Dealing with data, 331(6018), 639–806.

Senellart, P. P., & Blondel, V. D. (2003). Automatic discovery of similar words, Survey of Text Mining: Clustering, classification, and retrieval (M. Berry, ed.).

Sheng, J., Amankwah-Amoah, J., & Wang, X. (2017). A multidisciplinary perspective of big data in management research. *International Journal of Production Economics*, *191*, 97-112.

Shin, J., Oh, J., & Jeong, D. Y. (2020). An empirical study on the social marketing of companies in india. *Journal of System and Management Sciences*, 10(4), 86–101.

Su Jung Kim , Ewa Maslowska & Edward C. Malthouse (2017). Understanding the effects of different review features on purchase probability. *International Journal of Advertising*

Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, 58(4), 696-707.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, *36*(7), 10760-10773.

Thomson, R., Lebiere, C., & Bennati, S. (2014, April). Human, model and machine: a complementary approach to big data. In *Proceedings of the 2014 Workshop on Human Centered Big Data Research*, 27-31.

Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, 31(2), 198-215.

Wei Chen, Bin Gu, Qiang Ye, Kevin Xiaoguo Zhu (2019) Measuring and Managing the Externality of Managerial Responses to Online Customer Reviews. *Information Systems Research*

Xiao-Long, W. Y. Z. J., & CHENG, X. Q. (2013). Network Big Data: Present and Future [J]. *Chinese Journal of Computers*, *6*.

Yin, S., Wang, G., Qiu, Y., & Zhang, W. (2007, October). Research and implement of classification algorithm on web text mining. In *Third International Conference on Semantics, Knowledge and Grid (SKG 2007)*, 446-449.

Youssra, R., & Sara, R. (2018). Big data and big data analytics: concepts, types and technologies. *Int J Res Eng*, *5*(9), 524-528.

Zhang, K.Z.K., S.J. Zhao, C.M.K. Cheung, and M.K.O. Lee. 2014. Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Sy*stems, 67, 78–89.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, 74(2), 133-148.

Zou, K., & Wang, D. (2021). A study on consumer empathic response to advertising expressions: Focusing on mobile storytelling video advertising. *Journal of System and Management Sciences*, 11(1), 1-20.